Statistical System

Word Suggestion Problem:

1. What is character ngram?, context : word level ngram.

2. Dense vectorizer??, vectorizier with removed zeros.

3.

Randomness

----

coin(fair) => Flip 50% H, 50% T

biased coin => 70% H, 30% T

Dictionary vectorizer

======================

document => words [white space delimited]

doc against word_index(frequency) vector table

	1	2	3	4 ...
d1	freq or tf/idf			
d2				
d3				

Replace document to word like w1, w2, ... wn

Index: we create using ngrams of character.

How do we create ngram of characters:

n = 2

availabilities:

{av, va, ai, ...}

n = 3

{ava, vai, ial, ...}

consideration:

Multigrams, n = 2, 3, 4

All vectors:

V1 to Vn.

{av, va, ai, ..., ava, vai, ila, ..., avai, ...}

like bag of words for a document.

we index it and prepare a table.

================================

Query part:

Given word: avilabiltes

Vectorize: with same strategies. for example n = 2, 3, 4.

Vq

how find it?

Vq dot product with V1 to Vn.

v1 to vn and we sort based score and get first n words as high ranged words.

===================================

Next Meeting:

============

10th July, 2:00PM to 3:00PM